EXHIBIT F

US006119124A

United States Patent [19]

Broder et al.

[11] **Patent Number: 6,119,124**

[45] **Date of Patent: Sep. 12, 2000**

[54] **METHOD FOR CLUSTERING CLOSELY RESEMBLING DATA OBJECTS**

[75] Inventors: **Andrei Z. Broder**, Menlo Park; **Steven C. Glassman**, Mountain View; **Charles G. Nelson**, Palo Alto; **Mark S. Manasse**, San Francisco; **Geoffrey G. Zweig**, Oakland, all of Calif.

[73] Assignee: **Digital Equipment Corporation**, Maynard, Mass.

[21] Appl. No.: **09/048,653**

[22] Filed: **Mar. 26, 1998**

[51] **Int. Cl.**[7] **G06F 17/30**

[52] **U.S. Cl.** **707/103**; 707/3; 707/2; 707/5

[58] **Field of Search** 707/3, 2, 5, 103

[56] **References Cited**

U.S. PATENT DOCUMENTS

5,488,725	1/1996	Turtle	395/600
5,675,819	10/1997	Schuetze	395/760
5,724,571	3/1998	Woods	707/5
5,819,258	10/1998	Vaithyanathan	707/2
5,857,179	1/1999	Vaithyanathan	707/2
5,909,677	6/1999	Broder	707/3
5,937,084	8/1999	Crabtree	382/137

OTHER PUBLICATIONS

Brin et al.; Copy Detection Mechanisms for Digital Documents; Department of Computer Science; www.db.stanford-.edu/~sergey/copy.html.

Broder; Some Applications of Rabin's fingerprinting method; Methods in Communications, Security, and Computer Science; pp. 1–10; 1993.

Carter et al.; Universal Classes of Hash Functions; Journal of Computer and System Science; vol. 18; pp. 143–154; 1979.

Heintze et al.; Scalable Document Fingerprinting (Extended Abstract) found @ www.cs.cmu.edu/afs/cs/user/nch/www/koala/main.htm on Sep. 1997.

Karp et al.; The Bit Vector Intersection Problem; Proceedings 36th Annual Symposium of Computer Science, IEEE Computer Society Press, Oct. 23–25, 1995; pp. 621–634.

Shivakumar et al; Building a Scalable and Accurate Copy Detection Mechanism; Proceedings of 1st ACM Conference on Digital Libraries (DL'96), 1996.

Shivakumar et al; SCAM: A Copy Detection Mechanism for Digital Documents; Proceedings of 2ND International Conference in Theory and Practice of Digital Libraries; 1995.

Primary Examiner—Wayne Amsbury
Assistant Examiner—Mark Terry
Attorney, Agent, or Firm—Michael Buchenhorner

[57] **ABSTRACT**

A computer-implemented method determines the resemblance of data objects such as Web pages. Each data object is partitioned into a sequence of tokens. The tokens are grouped into overlapping sets of the tokens to form shingles. Each shingle is represented by a unique identification element encoded as a fingerprint. A minimum element from each of the images of the set of fingerprints associated with a document under each of a plurality of pseudo random permutations of the set of all fingerprints are selected to generate a sketch of each data object. The sketches characterize the resemblance of the data objects. The sketches can be further partitioned into a plurality of groups. Each group is fingerprinted to form a feature. Data objects that share more than a certain numbers of features are estimated to be nearly identical.

24 Claims, 8 Drawing Sheets